УДК 519.101

Н. И. БЕЗМЕНОВ, канд. техн. наук, ***С. В. КОВАЛЕНКО***,
Р. Б. МАНЧЕНКО, магистрант НТУ «ХПИ»,
В. Г. БОРИСОВ, канд. техн. наук

ОПРЕДЕЛЕНИЕ АВТОРСТВА ТЕКСТА С ИСПОЛЬЗОВАНИЕМ БУКВЕННОЙ И ГРАММАТИЧЕСКОЙ ИНФОРМАЦИИ

Метод, описаний у даній статті для визначення авторства тексту, ґрунтується на формальній математичній моделі зустрічі послідовності елементів тексту. Як елементи тексту використовуються послідовності букв і послідовності граматичних класів слів. Частоти вживання граматичних класів у тексті є досить стійкою характеристикою автора, їх можна використовувати, щоб вирішувати проблеми спірного авторства тексту.

Описанный в данной статье метод для установления авторства текста, основывается на формальной математической модели появления последовательности элементов текста. В качестве элементов текста используются последовательности букв и последовательности грамматических классов слов. Частоты употребления грамматических классов в тексте являются достаточно устойчивой характеристикой автора, их можно использовать, чтобы решить проблемы спорного авторства текста.

The method described in the given article for an establishment of authorship of the text, is based on formal mathematical model of occurrence of sequence of elements of the text. As elements of the text sequences of letters and sequences of grammatical classes of words are used. Frequencies of the use of grammatical classes in the text are enough the steady characteristic of the author, they can be used to solve problems of disputable authorship of the text.

Введение. В последние десятилетия наметилась тенденция поиска и выявления характерных структур авторского языка путем применения формально-количественных, статистических методов. Первые пробные шаги на этом пути предпринял еще в начале XIX века Н.А. Морозов. Интересно, что тогда же известный математик А.А. Марков выступил с критикой его подхода, суть которой состояла в том, что автор не произвел тщательной статистической проверки утверждений относительно устойчивости некоторых элементов авторского стиля (например, частицы «не»).

В настоящий момент среди широко известных спорных моментов определения авторства можно назвать активно обсуждавшийся в последние десятилетия спор – был ли написан роман-эпопея «Тихий Дон» М.А. Шолоховым или же был использован текст другого автора [1]?

Постановка задачи. Задача определения авторства ставится следующим образом. Пусть имеются достаточно длинные фрагменты прозаических произведений ряда авторов на русском или ином языке, использующем неиероглифическую письменность. О некотором анонимном фрагменте известно, что он принадлежит одному из этих авторов, но кому именно – неизвестно. Требуется установить его автора.

С определенной, достаточно высокой степенью вероятности это представляется возможным. Представленный метод базируется на учете статистики употребления пар элементов любой природы, идущих друг за другом в тексте (букв, морфем, словоформ и т.п.).

Предлагаемый метод основывается на формальной математической модели последовательности букв (или любых других элементов) текста [2]. По тем произведениям автора, которые достоверно им созданы, вычисляется матрица переходных частот употребления пар элементов (букв, грамматических классов слов и т.п.). Она служит оценкой матрицы вероятности перехода из элемента в элемент. Матрица переходных частот строится для каждого автора, при этом оценивается вероятность того, что именно данный автор написал анонимный текст (или фрагмент текста). Автором анонимного текста полагается тот, у которого вычисленная оценка вероятности больше [3].

Метод определения авторства. Пусть имеется W писателей, у каждого из которых есть N_w текстов, где $w = \overline{0,\ W-1}$. Введем такие обозначения:

$\{w\}$ – множество текстов автора w, $w = \overline{0,\ W-1}$

$\{n\}$ – множество всех текстов автора, которому принадлежит текст n;

$\Im$ – множество всех букв (для русского языка его мощность вместе с пробелами равна 32 – без буквы «ё»);

P_{ij}^{w} – вероятность перехода из буквы i в букву j в тексте автора w, $w = 0, ..., W-1$, $\boldsymbol{i \in \Im,\ j \in \Im}$.

Q_{ij}^{wn} – количество (частота) переходов из буквы i в букву j в каждом из текстов n автора w, $n \in \{w\}$, $w = 0, ..., W-1$, $i \in \Im$, $j \in \Im$.

Степень неопределенности в переходе от одной буквы к другой для некоторого автора k можно определить с помощью энтропии:

$$H_k = -\sum_{i \in \Im,\ j \in \Im} P_{ij}^{k} \ln P_{ij}^{k}\ , \qquad (1)$$

где P_{ij}^{k} – вероятность перехода из буквы i в букву j в текстах автора k.

Аналогичная характеристика при наличии информации о реализовавшемся переходе описывается следующей величиной [4]:

$$\tilde{H}_k = -\sum_{i \in \Im}\left(\sum_{j \in \Im} P_{ij}^{k}\right) \ln \sum_{j \in \Im} P_{ij}^{k} = -\sum_{i \in \Im,\ j \in \Im} P_{ij}^{k} \ln \sum_{j \in \Im} P_{ij}^{k}\ . \qquad (2)$$

Поэтому связи между буквами в текстах автора k можно оценивать величиной Λ_k, определяемой как разность величин H_k и $\tilde{H}_k$:

$$\Lambda_k = -\left(\sum_{i\in\Im,\ j\in\Im} P_{ij}^k \ln P_{ij}^k - \sum_{i\in\Im,\ j\in\Im} P_{ij}^k \ln \sum_{j\in\Im} P_{ij}^k \right) = -\sum_{i\in\Im,\ j\in\Im} P_{ij}^k \left(\ln P_{ij}^k - \ln P_{i\bullet}^k \right), \quad (3)$$

где $P_{i\bullet}^k = \sum_{j\in\Im} P_{ij}^k$.

Таким образом,

$$\Lambda_k = -\sum_{i\in\Im,\ j\in\Im} P_{ij}^k \ln \frac{P_{ij}^k}{P_{i\bullet}^k} . \quad (4)$$

Поскольку в реальных условиях в наличии могут быть только оценки вероятностей P_{ij}^k, можно говорить только об оценке величины Λ_k. Если обозначить через C^Σ общее количество переходов от одной буквы к другой во всех текстах множества Σ и воспользоваться для оценивания вероятностей всеми текстами $\{w\}$ каждого из авторов w, то оценкой максимального правдоподобия для Λ_w можно признать следующую величину:

$$L_w(\{w\}) = -\frac{1}{C^{\{w\}}} \sum_{i\in\Im,\ j\in\Im} Q_{ij}^{\{w\}} \ln \frac{Q_{ij}^{\{w\}}}{Q_{i\bullet}^{\{w\}}} , \quad (5)$$

где $C^{\{w\}}$ – общее количество переходов от одной буквы к другой во всех текстах множества $\{w\}$.

Исключая нулевые частоты, получаем:

$$L_w(\{w\}) = -\frac{1}{C^{\{w\}}} \sum_{i\in\Im: Q_{i\bullet}^{\{w\}}>0} \sum_{j\in\Im: Q_{ij}^{\{w\}}>0} Q_{ij}^{\{w\}} \ln \frac{Q_{ij}^{\{w\}}}{Q_{i\bullet}^{\{w\}}} . \quad (6)$$

Естественным является предположение, что исключение из рассмотрения любого из текстов множества $\{w\}$ не должно оказать существенное влияние на оценку для Λ_w. Если же во множество $\{w\}$ случайно попадет текст другого автора, то исключение этого текста из рассмотрения при вычислении оценки величины Λ_w должно сказаться на вычисленном значении.

Обозначим через $L_w(\{w\},n)$ оценку величины Λ_w по множеству $\{w\}$, из которого исключен текст n:

$$L_w(\{w\},n) = -\frac{1}{C^{\{w\}n}} \sum_{i\in\Im: Q_{i\bullet}^{\{w\}n}>0} \sum_{j\in\Im: Q_{ij}^{\{w\}n}>0} Q_{ij}^{\{w\}n} \ln \frac{Q_{ij}^{\{w\}n}}{Q_{i\bullet}^{\{w\}n}} , \quad (7)$$

где $Q_{ij}^{\{w\}n}$ – величина Q_{ij}^{wn}, вычисленная по всем текстам множества $\{w\}$ без текста n.

Таким образом, для любого текста n и любого автора $\{w\}$ должно выполняться следующее условие:

$$L_w(\{w\}) = L_w(\{w\},n), \text{ если } n \in \{w\}; \quad (8)$$

$$L_w(\{w\}) \neq L_w(\{w\},n), \text{ если } n \notin \{w\}. \quad (9)$$

Для установления авторства текста n, будем поочередно включать текст в каждое из множеств $\{w\}$, $w = \overline{0,\ W-1}$, и осуществлять ранжирование потенциальных авторов, определяя ранг $R_w(\{w\},n)$ каждого из них как отсчитываемый от нуля порядковый номер значения $|L_w(\{w\}) - L_w(\{w\},n)|$ в итоговой последовательности, получаемой после упорядочивания по возрастанию элементов последовательности

$$|L_w(\{w\}) - L_w(\{w\},n)|,\ w = \overline{0,\ W-1}. \quad (10)$$

Тогда автором текста предлагается считать следующего автора $\{\hat{w}\}$:

$$R_{\hat{w}}(\{\hat{w}\},n)w = \min_{w=\overline{0,W-1}} R_w(\{w\},n). \quad (11)$$

Развивая процедуру установления авторства, возможно выделение следующих единиц анализа [2]: *а)* пар букв в их естественных последовательностях в тексте (словах), а также пробелов между ними; *б)* пар букв в последовательностях букв в приведенных формах слов; *в)* пар наиболее обобщенных грамматических классов слов, частей речи, их последовательностей в предложениях текста (существительные, глаголы, прилагательные и т.п. *г)* пар менее обобщенных грамматических классов слов (а именно, таких семантико-грамматических разрядов, как одушевленные существительные, неодушевленные существительные и т.п.).

Применение алгоритмов сжатия данных в задаче определения авторства. Информационное расстояние между текстами можно посчитать с помощью программ сжатия, например *zip*, *rar* и др. Сжатый файл – это набор инструкций для разжимающей программы, позволяющий без потерь восстановить исходный текст. Хотя у разных пакетов набор инструкций в файлах разный, размеры файлов получаются приблизительно одинаковыми. Этот удивительный факт связан с тем, что современные компрессоры достигают почти максимальной степени сжатия текстов, оставляя минимальное количество информации, необходимой для того, чтобы буквально воспроизвести текст. Это минимальное количество называется сложностью по Колмогорову.

Чтобы определить сложность текста S относительно текста T нужно подклеить текст S к концу текста T и посмотреть, насколько хорошо сжимается эта добавка: $K(S|T) \approx K(\mathrm{TS}) - K(T)$. Выполнив сжатие текстовой информации, по полученному набору символов в дальнейшем, также возможно проведение анализа, направленного на определение авторства текста.

Выводы. Основным результатом проведенного исследования является то, что использование грамматической информации при решении задачи определения действительного автора текста является не только осмысленным, но и достаточно эффективным, а в некоторых случаях сопоставимым с использованием информации о встречаемости пар букв в тексте.

Интересен тот факт, что использование такой, простой единицы, как пара подряд идущих в тексте букв, дает более точные результаты, чем использование таких языковых категорий, как одиночные грамматические классы слов и их пары. Вполне возможно, что в буквенных парных структурах в преобразованном и, конечно, в неполном виде отображаются полные структуры морфем словоформ текста – префиксальные, корневые, суффиксальные и флективные. Тем самым, довольно большой объем словоизменительной и словообразовательной информации о структуре слов оказывается отображенным в статистике парной встречаемости букв, что и определяет довольно высокий уровень эффективности использования этой статистики для определения авторства текста.

Таким образом, подсчет частот употребления пар букв позволяет в некотором виде учесть информацию о словаре, который используется автором и, косвенно, информацию о предпочитаемых им грамматических конструкциях. Несмотря на то, что различия в частотах употребления конкретных пар букв, скорее всего, несущественны, поскольку сходятся к частотам, средним по языку, при увеличении объема текстов, «правдоподобие», учитывающее «общий» эффект изменения употреблений пар букв позволяет все же с высокой степенью точности определить истинного автора произведения.

Список литературы: ***1.*** *Хмелёв Д.В.* Распознавание автора текста с использованием цепей А.А. Маркова // Вестн. МГУ. Сер. 9: Филология. – 2000, № 02. – С.115-126. ***2.*** *От Нестора до Фонвизина*. Новые методы определения авторства. – М.: Издат. группа «Прогресс», 1994. ***3.*** *Фоменко В.П., Фоменко Т.Г.* Авторский инвариант русских литературных текстов // Методы количественного анализа текстов нарративных источников. – М.: Ин-т истории СССР. – 1983. – С. 86-109. ***4.*** *Ивченко Г.И., Медведев Ю.И.* Математическая статистика. – М.: Высш. шк., 1992.

Поступила в редколлегию 15.11.07